On the Governance of Semantic Artefacts in Dataspaces

Lucía Sánchez-González[1], Ana Iglesias-Molina[1], Oscar Corcho[1] and María Poveda-Villalón[1]

[1] *Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain*

Abstract

There is widespread agreement that the use of Semantic Artefacts (SA) (vocabularies, ontologies, etc.) in dataspaces is key for ensuring the interoperability of data, hence facilitating its integration. However, the different methodologies and ad-hoc practices for developing SA may incur in producing resources that do not meet the dataspace requirements, or that are not interoperable with the rest of the environment. The introduction of SA in dataspaces opens the door for SA governance to harmonize the efforts and resources within. This paper provides an overview of governance models for SA, along with the dataspace initiatives that address this concept, from which we extract the challenges to face for the implementation of SA governance in dataspaces.

Keywords

Dataspaces, Semantic Artefacts, Data Governance, Semantic Artefacts Governance

1. Introduction

Dataspaces were first conceived almost two decades ago as a data co-existence approach [1]. Most of the relevant initiatives have arisen more recently, such as the International Data Spaces Association (IDSA)[1]; the Gaia-X European Association for Data and Cloud [2]; or the Data Spaces Business Alliance (DSBA)[3], formed by the Big Data Value Association[4], FIWARE Foundation[5], Gaia-X and IDSA. Indeed, dataspaces have had such an impact that even the European Union is allocating significant resources for their implementation at European level [2].

The need for data harmonization and interoperability promotion between dataspace components has increasingly fostered the use of Semantic Artefacts (SA). The term Semantic Artefact refers to ontologies, terminologies, taxonomies, thesauri, vocabularies, metadata schemas, and other standards [3]. The European Commission itself emphasizes the importance in dataspaces

The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, May 26 – 27, 2024, Hersonissos, Greece

lu.sanchez@upm.es (L. Sánchez-González); ana.iglesiasm@upm.es (A. Iglesias-Molina); oscar.corcho@upm.es (O. Corcho); m.poveda@upm.es (M. Poveda-Villalón)

0000-0002-1685-4700 (L. Sánchez-González); 0000-0001-5375-8024 (A. Iglesias-Molina); 0000-0002-9260-0753 (O. Corcho); 0000-0003-3587-0367 (M. Poveda-Villalón)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://internationaldataspaces.org/

[2] https://gaia-x.eu/

[3] https://data-spaces-business-alliance.eu/

[4] https://bdva.eu/

[5] https://www.fiware.org/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

of domain-specific vocabularies and ontologies for the integration of data from heterogeneous data sources [2]. The IDS Information Model [4] is a clear example of how SA can achieve successful interoperability between dataspace components and roles.

As a result, the volume of SA available in dataspaces is growing rapidly. However, ontologies and vocabularies are most commonly developed following ad-hoc practices, which often result in resources that are not necessarily interoperable, reusable, and may even be obsolete, hindering their consumption and exploitation. This is where the concept of governance comes in. Governance can be defined as the set of political, institutional and administrative principles, rules, practices and processes through which and how decisions are taken and implemented [5]. While data governance has been defined as one of the essential elements of a dataspace [6], the concept of Semantic Artefact Governance (SAG) has been barely mentioned when it comes to dataspace components.

Given the importance of SA and the role of SAG for their correct development and management within dataspaces, this article reviews state of the art SAG frameworks and initiatives of dataspaces that address the concept of SAG. From this analysis, we identify the main challenges for implementing SAG in dataspaces, in order to promote governance to get the most out of dataspaces empowered with SA. The remainder of this article is structured as follows: We first present current proposals for SAG in Section 2. Then, we proceed to describe the dataspaces that implement or mention SAG in Section 3. Next, we identify the challenges of implementing SAG in dataspaces in Section 4. Finally, we draw the conclusions and future steps in Section 5.

2. Frameworks for Semantic Artefacts Governance

A search among major communities that develop, use and publish SA led to the identification of six relevant governance frameworks. We consider as SAG frameworks the initiatives that fulfill the following criteria: i) They must be focused on a network of SA, ii) there must be a community supporting them, and iii) they must define a set of guidelines (or principles, requirements, etc.) whose objective is to harmonize the development and publication of SA.

We provide a brief description of the identified frameworks, along with a comparison between them based on a series of features.

OBO Foundry [7]. The Open Biomedical Ontologies (OBO) consortium launched this initiative aiming for governing the increasing heterogeneity of ontologies in the biological domain. It comprises a set of detailed principles[6] that indicate how to develop ontologies, each one including recommendations, requirements and implementation guidelines. This proposal focuses on developing a family of interoperable ontologies within their community.
IOF [8]. The Industrial Ontology Foundry (IOF) revolves around creating interoperable ontologies for the digital manufacturing industry. It proposes, similarly to OBO, a set of governing principles[7], along with standards, tools and training materials (on purchase).
SAREF Publication Framework [9]. The European Telecommunications Standards Institute

[6] https://obofoundry.org/principles/fp-000-summary.html

[7] https://oagi.org/pages/technical-principles

Table 1
Summary of features of governance frameworks: principles (F1), guidelines, best practices of methodologies (F2), standards or requirements (F3), roles and responsibilities (F4), tutorials or training support (F5), quality assurance methods (F6), tooling support (F7) and scope (F8).

	F1	F2	F3	F4	F5	F6	F7	F8
OBO	✓	✓	✓	×	✓	✓	✓	Biology
IOF	✓	✓	✓	×	✓	×	✓	Manufacturing
SAREF	✓	✓	✓	✓	×	×	✓	IoT
GOMO	✓	✓	✓	✓	✓	✓	×	Multidomain
CROP	×	✓	✓	✓	×	✓	✓	Crops
IVOA	×	✓	✓	×	×	×	✓	Astronomy

(ETSI)[8] released a framework applicable for the SAREF[9] suite of ontologies [10], which provide the means for semantic interoperability in the Internet of Things (IoT) sector. This framework includes generic principles[10], actors, use cases, technical requirements for each step of the ontology's life cycle, and a set of best practices regarding naming conventions, metadata and ontology reuse.

GOMO [11]. The chemicals company BASF started a transition towards adopting semantic technologies to improve internal data management and exploitation. The Governance Operational Model for Ontologies (GOMO) was developed to establish how ontologies should be created and maintained within the company. This framework defines a set of governing principles, standards, best practices, and training materials.

CROP Ontology Governance and Stewardship Framework [12]. This framework was proposed by CGIAR (Consultative Group for International Agricultural Research) for the CROP Ontology Project, which collects crop-related ontologies [13]. It primarily focuses on defining the roles and corresponding responsibilities of the actors involved in the ontology design, development and maintenance. It also provides a set of additional guidelines [14] for data annotation, and quality assurance methods.

IVOA [15]. The International Virtual Observatory Alliance (IVOA) released a recommendation for the Virtual Observatory vocabularies to enhance the correct functioning and interoperability within the astronomical community. It provides guidelines for vocabulary content, metadata, management and publication, along with compliant tooling for enabling interoperability.

Table 1 summarizes the similarities and differences of the governance frameworks presented. We characterize them according to a set of features, which are described below. These features are extracted from the most common and shared elements that the analyzed governance frameworks define.

- **F1** – **Principles**. Set of fundamental rules based on the scope of the community that develops the SA. They establish the basis for the design of the rest of the components. For example, the *OBO Principle 1* states that "The ontology MUST be openly available"[11].

[8] https://www.etsi.org/
[9] https://saref.etsi.org/
[10] https://saref.etsi.org/principles.html
[11] https://obofoundry.org/principles/fp-001-open.html

F2 – **Guidelines, Best Practices or Methodologies**. Group of recommended procedures, processes and activities that explain how to develop the resource based on the principles and requirements established in the first place (e.g. Linked Open Terms (LOT) methodology [16]).

F3 – **Standards or Requirements**. Formal specifications and conditions created to satisfy the needs of the organization. For example, the SAREF Publication Framework in *Section 8.2.3 Requirements for usability and referencing* requires that "Each ontology module version should be available at least in Turtle, RDF/XML, and HTML formats"[9].

F4 – **Roles and Responsibilities**. List of actors involved in each of the phases of the SA life cycle and their corresponding activities (e.g., ontology engineer, domain expert, ontology maintainer).

F5 – **Tutorials or Training support**. Series of dissemination and teaching activities about the SAG framework. For instance, GOMO included as a fundamental pillar the organization of training workshop sessions to teach practitioners within BASF to develop ontologies according to the governance framework [11].

F6 – **Quality Assurance Methods**. Set of methods, metrics and procedures that allow evaluating the correct implementation of the standards, requirements or best practices. For example, the CROP Ontology Governance Framework provides a quality assessment workflow for ontologies that are candidates to be submitted to the network, to ensure that they follow the required guidelines.

F7 – **Tooling Support**. Specification of a series of tools to use for SA development and management. For instance, OBO Foundry has developed the ROBOT tool, which supports automation of ontology development tasks, focusing on OBO conventions [17].

F8 – **Scope**. Refers to the domain area of the SA. For example, the SAREF SAG refers to ontologies that model knowledge in the area of the Internet of Things.

3. Initiatives for the Governance of Semantic Artefacts in Dataspaces

In order to determine the status of the use of SAG in dataspaces, a search was carried out in official documents and technical reports on dataspace design and implementation initiatives. We consider those initiatives that explicitly mention SAG, or those that indicate the use of guidelines, tools or resources for SA management. This analysis of the main activities in dataspaces led us to the identification of two main initiatives.

IDS-Vocabulary Hub and Vocabulary Provider. In the IDS Reference Architecture Model (IDS-RAM[12]), the IDSA refers to two main elements related to SA governance: (i) The Vocabulary Hub, and (ii) the Vocabulary-related roles. The *Vocabulary Hub* serves as a platform to store, maintain, and publish shared vocabularies and related schema documents. In addition, this platform is thought to provide dataspace users with a series of tools that allow the creation, improvement, and publication of the terms of the vocabularies. However, it is specified that,

[12] https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4/

while it is required that these vocabularies employ RDF, it is not enforced the use of formal ontologies. Regarding *Vocabulary-related roles*, the IDS-RAM also defines a set of roles related to the management of vocabularies. For instance, in the documentation they mention roles such as Vocabulary creator, Vocabulary owner, Vocabulary publisher, Vocabulary consumer or Vocabulary user.

EOSC Interoperability Framework and the European dataspaces. EOSC (European Open Science Cloud)[13] has been defined as a key element to develop a science, research and innovation dataspace, and as a support for the implementation of the rest of sector specific dataspaces in Europe [18]. Among its multiple initiatives, the EOSC Interoperability Task Force published in 2021 the guidelines and principles that should drive the development of the EOSC Interoperability Framework [19]. Specifically, the document remarks the need for a semantic interoperability layer, where they address the use of SA to homogenize the interpretation and treatment of the exchanged data and all of its associated resources. They point out the lack of common and well-documented SA between communities, which is also affected by the absence of common and reference repositories. As a solution, they highlight the need for principles-based approaches and tools for the creation, maintenance, governance and use of SA. In addition, it is stated that EOSC should provide support for the maintenance of a repository of these SA, and a governance framework for such a repository. The establishment of these SAG initiatives by EOSC will be key to later being able to implement them in the SA used in the European dataspaces.

4. Challenges

The analysis carried out in Section 2 and Section 3, together with our expertise in dataspaces, SA and governance led us to identify the following challenges for implementing SAG in dataspaces.

Challenge I – Adoption of SAG by dataspaces. SA are gaining importance in dataspaces, as they have proven beneficial for enhancing interoperability and heterogeneous data integration[20]. As a result, they have already been incorporated in several initiatives [4], and more initiatives will follow in the near future [20]. While data governance in dataspaces is always considered, only a few initiatives mention how to govern their SA. Their proposed strategies to manage SA are too limited in comparison with SAG approaches outside dataspaces. Therefore, it is necessary to increase efforts to raise awareness among the actors involved in dataspaces about the fundamental role of the governance of these resources.

Challenge II – Generalization of SAG. Current SAG frameworks were designed to harmonize the ontology development efforts within a certain scope. In other words, they are limited to a specific domain of knowledge (e.g. OBO to biomedical ontologies) or purpose (e.g. GOMO for SA within BASF), providing ad-hoc practices for their needs (e.g. using OBO vs OWL2 for ontology implementation) that are difficult to map to other SAGs. In addition, all frameworks fail to provide all features identified in Section 2 and there is no agreement on the terminology used for each of their elements (i.e, one framework may use *principle* while another one may employ *good practices*). Lastly, to the best of our knowledge, none of them includes the governance

[13] https://eosc.eu/

of other semantic resources (e.g. queries, validation shapes, mappings). Hence, there is no holistic SAG framework that can be instantiated for each scenario and that includes all semantic resources that may play a role in dataspaces.

Challenge III – Coordination and limits between SAG and dataspace governance. Within dataspaces, specific governance frameworks are being designed [21]. The inclusion of SAG within these frameworks may be different from how current SAG frameworks are designed, specifically regarding the following aspects: (i) how the SAG framework overlaps with the dataspace governance framework; (ii) whether the SAG can be applied at dataspace level or stakeholder/users level, or both; and (iii) which specific parts of a SAG play a more relevant role, e.g. the key elements for interoperability and maintainability. The answer to these questions vary depending on the combination of dataspace and SAG framework.

5. Conclusions and Future Steps

With this work, we look into how SA are governed and their relevance for dataspaces. To this end, we perform a two-fold analysis. On the one hand, we identify and compare different SAG frameworks, with the aim of understanding what comprises a SAG framework; and assessing the implications of SAGs in real-world scenarios. From this analysis we observe that (i) SAGs are mostly domain- and/or community-specific, and (ii) there is no agreement regarding which elements should compose a SAG framework. On the other hand, we investigate which dataspace initiatives address the concept of SAG. Only two initiatives are identified, despite how emphasized and extended the use of SA in dataspaces is. Between them, only one actually defines some of the elements found in SAGs, while the other only mentions the relevance of implementing them in dataspaces.

The results of this analysis arise the following challenges, which define the future lines of research. First, a greater dissemination of the SAG concept is necessary among the dataspace community. This heavily relies on the second challenge: so far, there is no standard or reference SAG framework that allows for a clear identification and definition of the elements or features that a SAG framework should have. Among the initiatives that we analyze, we find discrepancies in the definition of each element, making their alignment difficult, thus hindering their implementation in dataspaces. Therefore, there is a need for identifying the different governance need scenarios for the development and maintenance of SA in dataspaces. Based on these scenarios, the elements that comprise the SAG frameworks to be used can be designed and adapted accordingly. This will open the door for addressing the last issue, on how to escalate and coordinate SAGs with dataspace governance.

In future steps, we plan to perform a more fine-grained analysis of the SAG models applied to dataspaces to elucidate the design of an abstract model suitable for different application scenarios. We will base the design and subsequent evaluation in close collaboration with dataspace initiatives, such as the Public Procurement Data Space (PPDS)[14], the Urban Data Space for the Green Deal (USAGE)[15], and INESData[16].

[14] https://europa.eu/!qx9WxQ

[15] https://www.usage-project.eu/

[16] https://inesdata-project.eu/content/en/index.html

Acknowledgments

Lucía Sánchez González is funded by the EU project USAGE - Urban Data Space for Green Deal (https://www.usage-project.eu/) which has received funding from the European Union's Horizon Europe Framework Programme for Research and Innovation under the Grant Agreement no 101059950 - call HORIZONCL6-2021-GOVERNANCE-01-17 (IA).

María Poveda-Villalón is funded by the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 101016854 (AURORAL).

References

[1] M. Franklin, A. Halevy, D. Maier, From databases to dataspaces: a new abstraction for information management, SIGMOD Rec. 34 (2005) 27–33. URL: https://doi.org/10.1145/1107499.1107502. doi:10.1145/1107499.1107502.

[2] E. Commission, Commission Staff Working Document on Common Data Spaces, Technical Report, 2024. URL: https://digital-strategy.ec.europa.eu/en/library/second-staff-working-document-data-spaces.

[3] C. Jonquet, J. Graybeal, S. Bouazzouni, M. Dorf, N. Fiore, X. Kechagioglou, T. Redmond, I. Rosati, A. Skrenchuk, J. L. Vendetti, M. Musen, Ontology repositories and semantic artefact catalogues with the ontoportal technology, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 38–58.

[4] S. Bader, J. Pullmann, C. Mader, S. Tramp, C. Quix, A. W. Müller, H. Akyürek, M. Böckmann, B. T. Imbusch, J. Lipp, S. Geisler, C. Lange, The international data spaces information model – an ontology for sovereign exchange of digital content, in: The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2020, p. 176–192. URL: https://doi.org/10.1007/978-3-030-62466-8_12. doi:10.1007/978-3-030-62466-8_12.

[5] J. Fritzenkötter, L. Hohoff, P. Pierri, S. Verhulst, A. Young, A. Zacharzewski, Governing the environment-related data space, SSRN Electron. J. (2022).

[6] E. Farrell, M. Minghini, A. Kotsev, J. Soler Garrido, B. Tapsall, M. Micheli, M. Posada Sanchez, S. Signorelli, A. Tartaro, C. J. Bernal, M. Vespe, M. Di Leo, B. Carballa Smichowski, R. Smith, S. Schade, K. Pogorzelska, L. Gabrielli, D. De Marchi, European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale, Scientific analysis or review KJ-NA-31-449-EN-N (online),KJ-NA-31-449-EN-C (print), Luxembourg (Luxembourg), 2023. doi:10.2760/400188.

[7] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al., The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nature biotechnology 25 (2007) 1251–1255. doi:10.1038/nbt1346.

[8] B. Kulvatunyou, E. Wallace, D. Kiritsis, B. Smith, C. Will, The industrial ontologies foundry proof-of-concept project, in: Advances in Production Management Systems.

Smart Manufacturing for Industry 4.0: IFIP WG 5.7 International Conference, APMS 2018, Seoul, Korea, August 26-30, 2018, Proceedings, Part II, Springer, 2018, pp. 402–409.

[9] SmartM2M, ETSI TR 103 608 V1.1.1: SmartM2M SAREF publication framework reinforcing the engagement of its community of users, Technical Report, ETSI, 650 Route des Lucioles, F-06921 Sophia Antipolis Cedex - FRANCE, 2019.

[10] R. García-Castro, M. Lefrançois, M. Poveda-Villalón, L. Daniele, The etsi saref ontology for smart applications: a long path of development and evolution, Energy Smart Appliances: Applications, Methodologies, and Challenges (2023) 183–215.

[11] A. Iglesias-Molina, J. A. Bernabé-Díaz, P. Deshmukh, P. Espinoza-Arias, A. Ayllón-Benítez, A. Fernández-Izquierdo, J. M. Ponce-Bernabé, S. Pérez, E. Ruckhaus, O. Corcho, J. L. Sánchez-Fernández, Ontology Management in an Industrial Environment: The BASF Governance Operational Model for Ontologies (GOMO), in: 2022 ISMB Bio-Ontologies Community of Special Interest, 2022. URL: https://oa.upm.es/74363/. doi:10.5281/zenodo.7007495.

[12] R. Mauleon, P. Jaiswal, L. Cooper, Winger, Borja, McNally, J. Detras, Michotey, C. Pommier, J. Pietragalla, Afolabi, Das, A. Rathore, I. Chaves, Makunde, Mendez, E. Salas, Hualle, E. Arnaud, Crop Ontology Governance and Stewardship Framework, Technical Report, CGIAR, 2022. doi:10.13140/RG.2.2.17471.48806.

[13] L. Matteis, P.-Y. Chibon, H. Espinosa, M. Skofic, H. Finkers, R. Bruskiewich, J. Hyman, E. Arnoud, Crop ontology: vocabulary for crop-related concepts 979 (2013).

[14] J. Pietragalla, L. Valette, R. Shrestha, M.-A. Laporte, T. Hazekamp, E. Arnaud, Guidelines for creating crop-specific Ontology to annotate phenotypic data: version 2.1. Alliance Bioversity International and CIAT, Technical Report, CGIAR, 2022. URL: https://hdl.handle.net/10568/110906.

[15] M. Demleitner, N. Gray, M. Taylor, Vocabularies in the VO Version 2.1, IVOA Recommendation 2023-02-06, Technical Report, Semantics Working Group, International Virtual Observatory (IVOA), 2023. https://www.ivoa.net/documents/Vocabularies/20230206.

[16] M. Poveda-Villalón, A. Fernández-Izquierdo, M. Fernández-López, R. García-Castro, Lot: An industrial oriented ontology engineering framework, Engineering Applications of Artificial Intelligence 111 (2022) 104755. URL: https://www.sciencedirect.com/science/article/pii/S0952197622000525. doi:https://doi.org/10.1016/j.engappai.2022.104755.

[17] R. C. Jackson, J. P. Balhoff, E. Douglass, N. L. Harris, C. J. Mungall, J. A. Overton, Robot: A tool for automating ontology workflows, BMC Bioinformatics 20 (2019). URL: https://www.osti.gov/biblio/1560605. doi:10.1186/s12859-019-3002-3.

[18] C. European Commission, Directorate-General for Communications Networks, Technology, A european strategy for data, communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions, 2020. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066.

[19] E. Commission, D.-G. for Research, Innovation, O. Corcho, M. Eriksson, K. Kurowski, M. Ojsteršek, C. Choirat, M. Sanden, F. Coppens, EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021. doi:doi/10.2777/620649.

[20] J. Theissen-Lipp, M. Kocher, C. Lange, S. Decker, A. Paulus, A. Pomp, E. Curry, Semantics in dataspaces: Origin and future directions, in: Companion Proceedings of the ACM

Web Conference 2023, WWW ’23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 1504–1507. URL: https://doi.org/10.1145/3543873.3587689. doi:10.1145/3543873.3587689.

[21] M. Dietrich, M. Gutierrez, D4.1: Phase 1 Governance Requirements and Endorsed Governance Scheme, Version 1.0, Green Deal Data Space and Foundation and its Community of Practice (GREAT), 2023. URL: https://www.greatproject.eu.